18/pats

Specification

DOCUMENT IMAGE PROCESSOR, ~~METHOD FOR~~ EXTRACTING DOCUMENT TITILE, AND METHOD FOR IMPARTING DOCUMENT TAG INFORMATION

Field Of The Invention

This invention relates to document image processor and document image processing method for storing and managing document images as image data, more specifically relates to the apparatus and the method for extracting title regions and marks attached by a user from a document image to use them as document tag information.

The Prior Art

Together with the increase of capability of data storage, document image processors rapidly become popular in which paper documents read from a scanner and etc. are stored and managed as document images that are image data.

It is arranged in such document image processor that each document image is registered corresponding to character strings that are document tag information like a keyword or a title so as to search a desired document image from plural document images stored in a data storage.

Fig. 19 shows the document tag information conceptually. As shown in the drawing, the document tag information, "Confidential" 191,

"A-company" 192, "Year 1999" 193 and "New car" 194, for example, acts as a keyword for a document image 190. Provided that a plurality of document tag information is attached to respective document image like this, it is possible to search the desired document image by narrowing down from those plural document tag information.

So far, a user has inputted by hand those document tag information at the storing of the document image. However, since it is a user that should input the document tag information, the more the number of documents increase, the larger the labor volume gets. Accordingly, such inputting operation is quite impractical. So in recent years the other apparatus has also appeared that permits to attach document tag information without assistance of hand labor by attaching a character string as document tag information after recognizing characters of the document image.

For instance, Japanese laid-open publication No. 8-147313 discloses a method of using a marked sheet. In the method, first, user marks a check box of document tag information to be attached to a document image, which is described on a marked sheet in a specific form. Then the marked sheet is read by a document image processor before the paper document is read, thereby the document tag information to be attached can be specified from the nominees of document tag information registered in advance. By the method, without using input device such as a keyboard or a pointing device, it is possible to attach the document tag information automatically to the document image to be registered.

Incidentally, it is very important for effective searching of document images to give appropriate document tag information to them. Specifically, it is a general searching method that specifies the document tag information corresponding to the desired document image from a list of plural document tag information displayed on a display. And to specify such document tag information quickly at the searching, respective document tag information should express the contents of the document directly.

The Japanese Laid-open Publication No. 8-202859 proposes the other method wherein a region to which a title-character string belongs (which is called a "title region" hereafter) is extracted from a document image, and the characters are recognized in the image of the title region, and then as a result of the recognition a title-character string gets to be document tag information. Since the title-character string represents the contents of the document directly, an image data processor adopting the title-region extracting can quickly specify the document tag information corresponding to the desired document image.

In the above method of extracting the title region disclosed in Japanese laid-open publication No. 8-202859, according to the aspect that the title characters are in the largest size among all of characters included in the document image, it is arranged that, after dividing the document image into plural regions (a region fitting the consecutive character rectangles together) and calculating the average character size in the respective regions, the region with the maximum of the average character size is extracted as a title region. Accordingly, it is natural that the

number of the title region extracted by the title-region extracting method is only one for a document image.

However, if there are plural documents with the approximate contents, titles of documents get approximate each other. Therefore, the conventional title-region extracting method had a problem that, when there are plural documents with approximate contents, it is impossible to specify the document tag information corresponding to the desired document image quickly.

In order to avoid the above problem it may be arranged at the preparing of a paper document that titles of similar contents should not be attached. However, it is not preferable to request a user to do the preparing operation.

On the other hand, the method in Japanese laid-open publication No. 8-147313 using a mark sheet has a very troublesome work that it is necessary to define the format of the mark sheet describing all items of the document tag information and to define the reading method of the mark sheet when a document image processor is configured regarding the software. Besides, in case of adding and registering nominees of new document tag information later on, the items of document tag information have changed. Thereby, it requests to reconstruct the format of the mark sheet and the reading method.

In addition, in case of using the mark sheet, it is hard for a user to visually confirm which the document tag information is attached, and there is a trouble that the inputting mistakes generates frequently.

The invention is proposed taking the above problems into consideration, and has an object to provide the document image processor for extracting title regions and marks attached by the user from a document image to use them as document tag information, and the method for extracting document titles, and the method for imparting document tag information.

SUMMARY OF THE INVENTION

The invention adopts the following means in order to achieve the objects.

First of all, a document image processor, as shown in Fig. 1, comprising region dividing means 103 for dividing a document image into a plurality of regions, and title-region extracting means 104 for extracting title regions from the entire regions according to a region average character size calculated per region divided by the region dividing means 103 adopts the following means.

First of all, after calculating a total average character size equivalent to an average height of characters in the entire regions, the title-region extracting means 104 compares the region average character size and an extracting criterion that is the total average character size multiplied by an extracting parameter, and then extracts as a title region regions with the region average character size larger than the extracting criterion. Accordingly, if the region has the region average character size larger than the extracting criterion, the region is extracted as a title region. Therefore, it is possible to extract a plurality of title regions from a document image.

In addition, the title-region extracting means 104 may be arranged to calculate the extracting criterions on a plurality of levels by using extract parameters on a plurality of levels. Thereby, the extracting judgment can be performed based on respective extracting criterions on a plurality of levels, so that it is possible to extract not only title regions but also subtitle regions (a region including a subtitle-character string composed of characters a little smaller than the size of the title character).

Further more, the title-region extracting means 104 may determine extracting parameters on a plurality of levels based on a maximum value of the region average character size divided by the total average character size. If the extracting parameter is calculated based on the maximum value of the region average character size instead of a fixed value, it is possible to obtain the extracting criterion more accurately.

And since the trim average method excluding characters larger than a specific rate and characters smaller than a specific rate is used at the time of calculating the total average character size and the region average character size, it is possible to improve the accuracy of the extracting further more.

Moreover, the image of characters included in the extracted title region can be converted to a title-character string of a character code string by character recognizing means 105. The correcting means 112 corrects the title-character string; thereby a user can change the title of the document image freely.

Secondary, in the document image processing for preparing and storing document images by reading a paper document, reference tag

information storage means 1215 is provided as shown in Fig. 12 for storing the reference tag information (nominees of document tag information) together with an attribute value of the reference tag information in advance.

Next, mark extracting means 1205 is provided for extracting specific marks attached on a paper document by a user. The mark indicates a general one that is imparted in order for a user to identify a paper document like a stamp, a seal, an illustration, a signature of specific handwriting, and etc.

Calculating means 120A is provided in order to calculate a characteristics value representing the characteristics of the mark according to the variance of pixels composing the extracted mark.

Document tag information imparting means 1208 is provided for comparing the attribute value and the characteristics value, selects the reference tag information with the maximum of similarity, and then imparts the selected reference tag information to the document image.

According to the above procedure, it is possible to automatically impart the document tag information to the document image based on the mark used in general at the document filing of user. Therefore, it is possible to perform the document management in office, and so on.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows a schematic functional block diagram of a document image processor in the embodiment 1 of this invention.

Fig. 2 shows a flowchart of the title-region extracting in the embodiment 1 of the invention.

Fig. 3 shows a flowchart of the title-region extracting in the embodiment 2 of the invention.

Fig. 4 shows a flowchart of the title-region extracting in the embodiment 3 of the invention.

Fig. 5 shows an explanatory diagram of the registration information management table in the embodiment 1.

Fig. 6 shows an explanatory diagram of the registration information management table in the embodiment 2.

Fig. 7 shows an explanatory diagram of the labeling.

Fig. 8 shows an explanatory diagram of the dividing of region.

Fig. 9 shows a diagram indicating the correlation among the height, the width, and the area of the character rectangle.

Fig. 10 shows a diagram representing the contents displayed on a display at the searching in the embodiment 1.

Fig. 11 shows a diagram representing the contents displayed on a display at the searching in the embodiment 2.

Fig. 12 shows a schematic functional block diagram of a document image processor in the fifth and sixth embodiments of the invention.

Fig. 13 shows an explanatory diagram of the registration image management table in the fifth and the sixth embodiments of the invention.

Fig. 14 shows an explanatory diagram of the mark management table in the fifth embodiment of the invention.

Fig. 15 shows an explanatory diagram of the reference tag information management table.

Fig. 16 shows an explanatory diagram regarding the extraction result of the mark image.

Fig. 17 shows an explanatory diagram the registration image management table in the sixth embodiment of the invention.

Fig. 18 shows an explanatory diagram of the mark management table in the sixth embodiment of the invention.

Fig. 19 shows an explanatory diagram expressing the conception of the document tag information.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

(EMBODIMENT 1)

The embodiments of the invention are explained hereafter referring to the drawings. The embodiments 1, 2, 3 and 4 are explaining an image data processor for extracting plural titles from a paper document.

Fig. 1 shows a schematic functional block diagram of a document image processor to which the present invention is applied. The configuration of the apparatus will be explained hereafter together with the procedure of the document image registration.

First, image inputting means 101 like a scanner, for example, performs the optical/electrical converting of a paper document, and then a document image 108a that is multi-levels image data is obtained. After image processing means 111a adapts the data to the storing (the compressing, for example), the document image is registered in a

document image area Aa of storage means 108. It may certainly be arranged that the multi-levels image data be registered in the document image area Aa without changing where the image processing means 111a is not provided.

The document image from the image inputting means 101 is inputted into not only the image processing means 111a but also image processing means 111b. Here, the document image is converted to binary image data and then stored into an image memory 107. While the image memory 107 stores the image data, character rectangle creating means 102 performs the following labeling referring to the document image stored in the image memory 107. The labeling is the processing, regarding black pixels among pixels to which a black pixel to be given attention (which is called a "target pixel" hereafter) is contiguous on the 8 directions, that is on the top side, on the upper right side, on the right side, on the lower right side, on the down side, on the lower left side, on the left side, and on the upper left side, for giving the black pixels the same labeling value (identification information) as that of the target pixel. That is to say, as shown in Fig. 7, where 8 pixels, W1, W2, W3, W4, W6, W7, W8 and W9 are contiguous with the target pixel W5, the character rectangle creating means 102 gives a labeling value same as that of the target pixel W5 to the black pixels W2, W3 and W8. According to such labeling, the same labeling value can be given per a black pixels contiguous component in the document image (per a group of black pixels).

Next, the character rectangle creating means 102 prepares a character rectangle by cutting off the black pixel contiguous component attached with the one same labeling value, and then transfers the character rectangle to region dividing means 103. Here, the "character rectangle" means a circumscribed rectangle of a black pixel contiguous component. However, there is a case where some characters are not always configured by one black pixel contiguous component. In consideration of this, it can be arranged that a section of the black pixel in the document image is expanded before the labeling. Specifically, it is the process for converting the 8 pixels contiguous to the target pixel to black pixels. The processing is repeated by appropriate times (generally twice or triple), thereby the section of black pixel is enlarged, and accordingly it is possible to combine respective black pixel contiguous components apart from each other within a character into one unit. If the labeling is performed after the above-mentioned processing, it is possible to prepare the character rectangle per character precisely.

When the character rectangle creating means 102 completes the processing, the region dividing means 103 detects areas adjacent to respective character rectangles, and then divides the document image to regions by combining the character rectangles contiguous with each other. For instance, the region dividing means 103 that received the character rectangles C1 to C12 as shown in Fig. 8 combines the character rectangles C1 to C4, C5 to C9, and C10 to C12 respectively and then divides the document image into regions E1, E2 and E3.

According to thus region dividing, the document image can be divided into regions per character string. The judgment whether the character rectangles are contiguous with each other or not, or whether there is a blank line between the character rectangles or not, should be determined by using proper threshold values of a character gap and a line space.

As a result of the above processing, it is possible to obtain the information of the total character size in the document image (which will be described later), the number of the divided regions, and the number of the character rectangles in each region. It is arranged in the invention that the serial number starting from 1 be given to each divided region and also be given to each character rectangle included in the region respectively. Hereinafter, the number of character rectangles in the n-th region is represented by $NumChar_n$, and the size of the m-th character in the n-th region is represented by $SizeChar_{n,m}$.

Incidentally, as shown in Fig. 9, even if the character font of the same point is adopted, the width W1 to W4 and the area A1 to A4 varies extremely depending on a kind of the character, conversely the height of a character rectangle varies a little. Therefore, the invention may adopt as the character size "the height of a character rectangle" that reflects the number of points of a character font comparatively correctly.

Here, title-region extracting means 104 extracts as a title region only specific regions from all regions divided as above. The title-region extracting is explained hereinafter according to a flowchart shown in Fig. 2.

Fist, the title-region extracting means 104 calculates a region average character size per region (Fig. 2, Step 1). The region average character size is an average value of size of all characters included in a region. The region average character size in the n-th region, $SizeReg_n$, is found to be the sum of all character size $SizeChar_{n,m}$ of all characters included in the region divided by the number of character $NumChar_n$ in the region. This correlation is represented by the following equation.

[Equation 1]

$$SizeReg_n = \Sigma SizeChar_{n,m} / NumChar_n$$

Next, according to the region average character size $SizeReg_n$ per region and the number of characters in the region $NumChar_n$ that are calculated as above, the total average character size $SizeALL$ in the document image is calculated by the following equation (Fig. 2, Step 2).

[Equation 2]

$$SizeALL = \Sigma (SizeReg_n \times NumChar_n) / \Sigma numChar_n$$

The method of calculating the region average character size $SizeReg_n$ and the total average character size $SizeAll$ is not restricted to the above method. For example, it is possible to adopt the Trim Average Method (a method of calculating the average after excluding a specific ratio, for example, 10% of data, from the minimum value side and the maximum value side), which will be explained later.

Here, according to the judgment whether the following equation of judging the extracting is established or not, the title-region extracting means 104 performs the extracting judgment of the title region.

[Equation 3]

$$SizeReg_n >= SizeALL \times \alpha$$

That is to say, after comparing the calculated total average character size *SizeALL* multiplied by a extracting parameter α (the extracting criterion) and the region average character size $SizeReg_n$ per region, the regions where the equation of the extracting judgment is established are extracted as a title region (Fig. 2, Step 3 to 4 to 5). The extracting parameter α should be a constant that is larger than 1.0, and it is preferable to be 1.2, for example.

When the extracting judgment is performed for all regions by repeating the above procedures (Fig. 2, "NO" in Step 3), the title-region extracting is completed. Then respective title-region images 108b extracted here are registered in a title area Ab of the storage means 108.

Next, character recognizing means 105 cuts the title-region images extracted from the document image off, performs the character recognizing for each title-region image, and then obtains title-character strings that are a character code string. Those title-character strings thus obtained are transferred to display control means 110 via correcting means 112, and are presented to a user by displaying them in a list view on a display that is not shown (see Fig. 10(I)).

The user confirms each title-region image and title-character string thus displayed, and if he wants to register one of the title-character strings in a same way as shown on the display, he instructs instruction inputting means 109 to register it. Then, the title-character string is transferred from the character recognizing means 105 to document registering means 106.

On the other hand, if the user wants to correct or change any of the title-character strings, he double-clicks a title-character string thus displayed by making use of a pointing device of the instruction inputting means 109. According to the double-clicking, the correcting means 112 instructs the display control means 110 to wink the title-character string on the display and display the cursor within the character string. Then the user, operating the keyboard of the instruction inputting means 109, inputs the corrected character string in the correcting means 112, thereby the character string after the cursor can be replaced with the corrected character string. The character string thus corrected is inputted to the character recognizing means 105 from the correcting means 112; thereby the title-character string is corrected. Likewise, when the user instructs to perform the registration by the instruction inputting means 109; the corrected title-character string is transferred from the character recognizing means 105 to the document registering means 106.

However, in case where the confirmation and the correcting are not programmed, the contents recognized by the character recognizing means 105 is to be transferred to the document registering means 106 as it is, without displaying it on the display.

After receiving the title-character string, the document registering means 106 registers the registration information composed of the storage pointer of the document image 108a and the title-region image 108b in the storage means 108, the title-character string, and the position and size of the title region in the document image into a registration information management table 108c formed in the table area

Ac on the storage means 108 (see Fig. 5). Here, the storage pointer of the document image 108a can be obtained from the document image area Aa on the storage means 108, the storage pointer of the title image 8b can be obtained from the title area Ab on the storage means 108, and the position and the size of the title region can be obtained from the character recognizing means 105.

After the registration information management table 108c is prepared as above, where the instruction inputting means 109 inputs the instruction of the searching of the document image, the display control means 110 displays in a list view the title-region images and the title-character string stored as above on the display (Fig. 10(I)).

When the user selects a title (a title-region image or a tile-character string) from the listed titles on the display by making use of the instruction inputting means 109, the display control means 110 displays on the display the document image corresponding to the title. At this time, as shown in Fig. 10(II), it is preferable that the title region in the document image is demonstrated by circumscribing it with a rectangular frame F. The rectangular frame F can be prepared according to the position and the size of the title region registered in the registration information management table 108c.

In addition to the above method of selecting either one from the list displayed on the display, it is needless to say that the method can be adopted that, when specific document tag information is inputted from the instruction inputting means 109, if the title corresponding to the

specific document tag information has been registered in the registration information management table 108c, the corresponding document image may be indicated on the display.

In accordance with this embodiment described above, since it is arranged that regions with the region average character size larger than the extracting criterion be extracted as a title region, it is possible to extract plural title regions from a document image. Therefore, even if there are many document images having similar contents, it is possible to quickly specify the document tag information (the title) corresponding to the desired document image.

The above explanation does not refer to the procedure when there is no region in which the extracting judgment equation is established at the title-region extracting. However, in this case, the intention that no title region can be extracted is displayed on the display and the input of a character string to be the document tag information is requested to the user. The user inputs the character string in response to the request, and then the inputted character string can be used as the title-character string for the document image.

(EMBODIMENT 2)

It is arranged in the embodiment 1 that, if the regions have the region average character size larger than the extracting criterion, those be extracted likewise as a title region regardless of the value of the region average character size. Therefore, it is impossible to perform the appropriate display processing based on the value of the region average character size, that is to say, the processing of displaying in a list view the

title-character strings without displaying in a list view the subtitle-character strings composed of characters smaller a little than the title character. In the embodiment of the invention, the above-mentioned problem is settled by arranging to calculate plural extracting criterions by using plural levels of extracting parameters and to extract the title regions corresponding to the level attributes (information indicating the level of the extracting). The configuration will be explained hereafter regarding to the points different from that of the embodiment 1.

The title-region extracting means 104, that calculated the region average character size $SizeReg_n$ and the total average character size *SizeAll* according to the same procedure as in the embodiment 1, performs the extracting judgment of plural levels according to the result whether the following equation for the extracting judgment on the plural levels is established or not.

[Equation 4]

$$SizeReg_n >= SizeALL \times \alpha_p$$

α_p in the above equation is a extracting parameter on the p-th level (the level p), and the value of α_p should be predetermined so as to satisfy the condition of the equation 5. When the extracting judgment is performed on 5 levels, it is preferable that each parameter should be determined as approximate α_1=1.5, α_2=1.3, α_3=1.2, α_4=1.15, and α_5=1.1.

[Equation 5]

$$1.0 < \alpha p < \ldots\ldots < \alpha_3 < \alpha_2 < \alpha_1$$

It is explained according to the flowchart shown in Fig. 3. The

title-region extracting means 104 performs the extracting judgment on every level in the order from the level 1 (Fig. 3, Step 14 to 15 to 14). When the extracting judgment equation is not established on either level, the region is not extracted as the title region but the title-region extracting means 104 performs the extracting judgment for the next region (Fig. 3, Step 14 to 13 to 14 to 15). On the other hand, when the extracting judgment equation is established on either one level, the region is extracted as the title region on that level (corresponding to the level attribute) and then the extracting judgment is performed for the next region (Fig. 3, Step 15 to 16 to 13 to 14 to 15).

After the extracting judgment is performed for every region by repeating the above procedure (Fig. 3, "NO" at Step 13), the title region extracting is completed.

Besides, when there is no region where the extracting judgment equation is established, the character string inputted by a user is used as the title-character string, which is the same as the embodiment 1. The level attribute of this title-character string is set as the level 1 and the total number of level is set as 1.

In addition, the extracted title-character string can be changed or corrected, which is the same as the embodiment 1.

Fig. 6 is an explanatory diagram of the registration information management table 1 in this embodiment. It is arranged that the "level attribute" field 601 and the "number of total levels" field 602 be added to the configuration described in the embodiment 1. And when there is any region extracted on the level 1 according to the extracting

judgment of the 5 levels, the document registering means 106 registers "5" on the "number of total levels" field 602 and "1" on the "level attribute" field 601 respectively.

Fig. 11 is a diagram showing the contents displayed on the display at the searching in this embodiment. It is arranged that each level attribute of the titles displayed in a list view on the upper portion can be selected by the instruction inputting means 109. And the display control means 110 displays on the display in a list view the titles within the scope selected as above, referring to the "level attribute" field 601 and the "number of total levels" field 602 in the registration information management table 108c.

In accordance with this embodiment described above, it is arranged that the extracting criterions of plural levels be calculated by using the extracting parameters of plural levels and the title regions be extracted corresponding to each level attribute. Therefore it is possible to perform various processing according to the value of the region average character size, for example, for displaying in a list view the title-character strings only, without displaying in a list view the subtitle-character strings.

(EMBODIMENT 3)

It is arranged in the embodiment 2 that the extracting parameters of plural levels be predetermined (as a fixed value), however, it is preferable that the extracting parameters should be determined according to respective characteristics of the inputted document images. It is arranged in this embodiment that the extracting

parameters of plural levels are determined according to the value equal to the maximum value of the region average character size divided by the total average character size (see Fig. 4, Step 23). The configuration will be explained regarding only the point different from that of the embodiment 2.

After calculating the region average character size $SizeReg_n$ and the total average character size *SizeAll* according to the same procedure as in the embodiment 2, the title-region extracting means 104 calculates first the value α_1 that is the maximum value of the region average character size, $max\{SizeReg_n\}$, divided by the total average character size *SizeAll*, according to the following equation.

[Equation6]

$$\alpha_1 = max\{SizeReg_n\} / SizeAll$$

Next, by using the following equation, the title-region extracting means 104 determines the extracting parameters α_p on each level in accordance with thus calculated α_1 and the total number of levels P (P>=1) for the extracting judgment.

[Equation 7]

$$\alpha_p = \alpha_1 - (p-1) \times (\alpha_1 - 1) / P$$

For instance, the extracting judgment is performed on 5 levels when α_1 is 1.5, the extracting parameters α_1 to α_5 on each level are calculated as follows.

[Equation 8]

$$\alpha 1 = 1.5-(1-1)\times(1.5-1) / 5 = 1.5$$

$$\alpha 2 = 1.5-(2-1)\times(1.5-1) / 5 = 1.4$$

$\alpha 3 = 1.5-(3-1)\times(1.5-1) / 5 = 1.3$

$\alpha 4 = 1.5-(4-1)\times(1.5-1) / 5 = 1.2$

$\alpha 5 = 1.5-(5-1)\times(1.5-1) / 5 = 1.1$

According to the equation 7, the extracting parameter α_p on each level can be determined so as to be equidistance between thus calculated $\alpha 1$ and 1.0.

The procedure after the above steps is the same as that of the embodiment 2 excluding the extracting judgment by using the extracting parameter determined as described above, which is not explained here.

However, the above method has an incompetent point that the region of text is extracted as the title region by mistake when any title region is not existed in the document image, this is because $\alpha 1$ is a value near to 1.0, for example, 1.03. In order to avoid such trouble, this invention is arranged that the extracting parameter under a specific value 1.05, for example, not be adopted.

In addition, when the difference between extracting parameters on each level is not more than a specific value, 0.03, for example, the extracting judgment cannot be performed precisely. Accordingly, it is arranged that the set value of the extracting parameter be corrected so that the difference between the extract parameters on each level may be said specific value (0.03). Practically, in the above case, 0.03 subtracted from α 1 sequentially is determined as the respective extracting parameters of each level.

As a result of the above, there is a possibility that the total number of levels P reduces. In this case, the real number of levels (which

is the reduced number of levels subtracted from the total number of levels P) is registered as the total number of levels P on the "number of total levels" field 602 in the registration information management table 108c.

In this embodiment as described above, it is arranged that the extracting parameters should not be fixed, but determined based on the characteristics of the inputted document image. Therefore it is possible to perform the extract determination precisely.

(EMBODIMENT 4)

In each of the above-mentioned embodiments, the characters of the title region in the relatively large size are also taken into the calculation of the total average character size, and the small characters such as a comma, a period, and punctuation are also taken into that calculation. Thereby the calculation result has an inclination to bring down the accuracy of the title extracting. Therefore it is arranged in this embodiment that the total average character size be calculated based on the total characters in the document image excluding the characters of which size is larger than a specific ratio (90%, for example) and characters of which size is smaller than a specific ratio (10%, for example). That is to say, the trim average method is adopted here. In addition, even when the region average character size is calculated, the same trouble occurs, too. Therefore, it is possible to apply the trim average method to the calculation of the region character size.

As a result, it is possible to calculate the total average character size and the region average character size regarding the characters

excluding a period, a comma and punctuation, so that the value of the title-region extracting can be precise more than ever.

Here, in each of the prescribed embodiments, the total average character size is calculated from the region average character size. But, when the method is applied to the trim average method, all characters included in the title region cannot be excluded for the calculation of the total average character size because the each region omits the large-sized characters and the small-sized characters. Therefore, it is arranged in this embodiment that the total average character size be calculated for the total characters in the document image once again.

However, even in case of using the trim average method, it is needless to say that either one of the specific value in the embodiment 1 or the level value in the embodiments 2 and 3 can be used as the extracting parameter.

The each explanation of the above embodiments does not refer to the number of the original image document, but the number of the original is not limited particularly. That is to say, even if there is only one sheet or plural sheets, it is possible to obtain the same effect if the same extracting parameter is used in each page. In particular, it is possible in the embodiments 2 and 3 to extract the title region and the sub-title region from one document composed of a plural page such as a data of the thesis by using the same extracting parameter in a plural page.

According to the above explanations, the height of the character rectangle is adopted as the character size, but the width or the area of the character rectangle can be adopted as the character size.

As illustrated in Fig. 1, since the image processing means 111a and 111b are provided at the stage before the storage means 108 and the image memory 107 respectively, it is possible to use a binary image as a document image for the title extracting while a compressed image or a multi-valued image as a document image data to be stored in the document image area Aa of the storage means 108. Thereby, it is possible to display in various way the document images obtained as a result of the searching based on the title regions thus extracted, like displaying in color.

(EMBODIMENT 5)

Here in the embodiments 5 and 6 is explained about the document image processor that recognizes the marks attached on the paper document as the document tag information.

First, marks such as a title or a keyword are given to any page composing the paper document by a user. Here the mark indicates a general mark given by a user so as to identify the paper document, like a stamp, a seal, an illustration, a signature of specific handwriting, and so on.

When the document image processor in the invention stores the paper document composed of a plurality of pages, it is need to judge which page of the paper document the marks are attached to. At this time, though there is a method for detecting the marks after the searching all over the total pages of the paper document, there is a problem in the method that it takes much time for the detecting.

The method to solve said problem is as follows; for instance, the document image processor is predetermined so as to perform the detecting

of the marks for the only first page.

In this embodiment of the invention, the marked pages (which is called "document tag information appointing page" hereafter) 21 and 24 can be distinguished from others by describing the specific 2-dimensional code image 26 on the specific position of the lower right, as shown in Fig. 13(b).

Fig. 12 shows a block diagram of the document image processor in the embodiment 5 of the invention. The procedure of the processing by the document image processor will be explained hereafter.

First, the image inputting means 1201 electronized the paper document by using an optical/electrical converter such as a scanner, a digital integrated apparatus, and so on, and then inputs them as the document images. Here, as shown in Fig. 13, the document tag information attached to the document tag information appointing page 21, "Confidential", "A-company", and "Year '99", should be given to the inputted images 22 and 23, and the document tag information attached to the document tag information appointing page 24, "Confidential" and "B-company", should be given to the inputted image 25. And the image inputting means 1201 is inputted the document tag information appointing page 21, the input image 22, the input image 23, the document tag information appointing page 24 and the input image 25, in those order.

The inputted document image is stored in the image memory 1202 temporarily, for which the image data compressing means 1203 performed the data compressing. After that, said data is stored in the image storage area 1211 of the storage means 1210. At this time, in order

to identify each document image thus stored, an image ID is given to the document image respectively. The image ID is registered in the "image ID" field 121 of the registration image management table 1212 shown in Fig. 13(a). In addition, the pointer information pointing to the image data stored in the image storage area 1211 of the storage means 1210 is registered in the "pointer to image data" field 122 of the registration image management table 1212.

On the other hand, the document image stored in the image memory 1202 is also sent to the mark extracting means 1205 after the binarization by the binarization means 1204. The mark extracting means 1205 judges whether the specific two-dimensional code image is at the predetermined position of the lower right of the image or not, and determines whether the inputted document image is the document tag information appointing page or not respectively.

At this time, if the document image is determined as the document tag information appointing page, "1" is registered on the "document tag information appointing page flag" field 123 of the registration image management table 1212, if not, "0" is registered on it. The flag is applied to the identification that the document image is the document tag information appointing page attached with the marks only and does not contains the content as the text of the paper document. For instance, after the document tag information are given to the document image according to the other-mentioned method, the document image corresponding to the document tag information appointing page can be deleted according to the flag. Thereby it is possible to avoid a waste of the

memory resources.

A mark management group No. is given to the entire document images inputted between the first document tag information appointing page and the next one. In addition, the mark management group No. is registered on the "mark management group No." field 125 of the registration image management table 1212. Here, it means that the document image to which the same document tag information is given is imparted the same mark management group No.

The next explanation refers to the processing that the mark extracting means 1205 extracts the marks from the document image determined as the document tag information appointing page according to the above processing.

First, as described in the embodiment 1 it labels the entire regions excluding the regions attached with the two-dimensional code among the document tag information appointing pages. Among a plurality of the black pixels contiguous components obtained by the labeling, the components that have the distance from each other smaller than the specific threshold value are combined to one region. The regions thus obtained are corresponding to the regions of marks 41 to 43 respectively, as shown in Fig. 16. Those regions are extracted, and thereby each mark image can be obtained.

The number of marks extracted from each document tag information appointing page is registered on the "number of marks" field 124 of the registration image management table 1212.

In addition, in order to manage the information of the extracted

mark images, each mark image is attached with a mark ID, and then registered on the "mark ID" field 131 of the mark management table 1213 as shown in Fig. 14. The mark management group No. of the document tag information appointing page attached with the mark is registered on the "mark management group No." field 132 of the mark management table 1213. Regarding each of the mark images extracted from the document tag information appointing page, the information about the position and the size (the width and the height) of the mark image within the document tag information appointing page are registered on the "position" field 134 and the "size" field 135 of the mark management table 1213 respectively.

In this embodiment, the document images inputted between the first document tag information appointing page and the next one is attached with the same mark management group No., and managed as a series of document images attendant on the first document tag information appointing page. It can be considered as another management method that only the specific document image inputted after the document tag information appointing page is given the mark management group No. and the other document images are not given any number. This method can be applied when a user wants to give the table of contents to the specific document image.

Next, the characteristics value calculating means 1206 of the calculating means 120A calculates the numerical value representing the characteristics of the mark image extracted by the mark extracting means 1205. The invention applies the characteristics value of the Moment Invariants of the well-known prior arts to this numerical value. The

following explanation is made regarding the Moment Invariants in brief.

When the coordinates of a pixel is represented by (i, j) and the value of the pixel is represented by I(i, j), I is a function that satisfies I=1 for the black pixel meanwhile satisfies I=0 for the white pixel. The m_{pq} defined by the Equation 9 is called the (p+q)-dimensional moment.

[Equation 9]

$$m_{pq} = \Sigma_i \Sigma_j i^p j^q I(i, j) \qquad p,q = 0, 1, 2, \ldots$$

In case of applying the above m_{pq}, the center of gravity (x, y) of the two-dimensional image is represented by the Equation 10.

[Equation 10]

$$x = m_{10} / m_{00}$$

$$y = m_{01} / m_{00}$$

μ_{pq} defined by the Equation 11 according to the center of gravity thus calculated is called the center moment.

[Equation 11]

$$\mu_{pq} = \Sigma_i \Sigma_j (i-x)^p (j-y)^q I(i, j)$$

The numerical values M1 to M6 calculated as follows by the Equation 12 according to the above center moment are defined as the characteristics value of the corresponding two-dimensional image (or on the Moment Invariants).

[Equation 12]

$$M1 = \mu_{20} + \mu_{02}$$

$$M2 = (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2$$

$$M3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2$$

$$M4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2$$

$$M5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2]$$
$$+ (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]$$
$$M6 = (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]$$
$$+ 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03})$$

Since those characteristics values are unchangeable even in case of the rotation or the translation of the two-dimensional image, those become the effective value for characterizing the mark like the embodiment of the invention when a user gives a specific mark to a paper by hand.

The characteristics value calculated by the characteristics value calculating means 1206 is given to the similarity calculating means 1207 of the calculating means 120A, where the similarity between these characteristics value and the attribute value of respective reference tag information is calculated. In order to explain this method, the following is the explanation about the management method of the reference tag information and the calculating method of attribute value of the respective reference tag information.

The reference tag information is the data correlated with the mark that a user will use in the future (which is called the "reference mark" hereafter), specifically, and the nominee of the document tag information like the character string playing a role as a keyword for the inputted image. The reference tag information is registered on the "reference tag information" field 141 of the reference tag information management table 1214 as shown in fig. 15(a). The image data of the reference mark is stored in the reference tag information storage means 1215. The pointer to this image data is registered on the "pointer to

reference mark image" field 142 of the reference tag information management table 1214. The characteristics value calculating means 1206 calculates six characteristics values of those reference marks on the Moment Invariants, which are registered on the "attribute value (M1 to M6)" field of the reference tag information management table 1214. That is to say, those characteristics values are the attribute values of the respective reference marks.

The distance between the attribute value of the reference mark thus calculated and the respective characteristics value on the Moment Invariants of the mark image extracted from the inputted image is calculated by the Equation 13.

[Equation 13]

$$L = (m1-M1)^2 + (m2-M2)^2 + (m3-M3)^2 + (m4-M4)^2 + (m5-M5)^2 + (m6-M6)^2$$

The above M1 to M6 represent the attribute value of the reference mark, while the above m1 to m6 represent the characteristics value of the extracted mark image. It expresses that the smaller the distance L calculated by the above equation, the higher the similarity of the extracted marked image and the reference tag information is.

The document tag information imparting means 1208 specifies a reference mark of which similarity is the maximum value, and selects the reference tag information of the reference mark as the document tag information of the inputted document image, and then imparts the information to the document image. In addition, the document tag information is registered on the "document tag information" field 133 of the

mark management table 1213.

By applying the above-mentioned processing, it is possible to automatically impart the document tag information to the inputted document image respectively. By using the information of each table obtained here, the searching of the image can be performed according to the following procedure.

First, when a user selects one of document tag information to be used for the searching, the mark management group Nos. corresponding to the document tag information can be specified from the mark management table 1213. Additionally, the image IDs of the document image attached with the mark management group No. and the pointer information to the document image data can be specified from the registration image management table 1212. The document image specified here becomes the image correlated with the document tag information designated by the user. By designating a plurality of document tag information, it is possible to narrow down the image data to be searched.

When the document tag information with the maximum of the similarity calculated by the similarity calculating means 1207 has the distance L from the extracted mark image that is larger than the predetermined threshold value, it is judged that there is no registered document tag information to be correlated with this mark image but new reference mark is inputted. In this case, the mark image is displayed according to the information of the "position" field 134 and the "size" field 135 of the mark management table 1213 and the "pointer to image data" field 122 of the registration image management table 1212, and then the

user is asked to register the document tag information to be correlated with the new reference mark.

The document tag information inputted here is newly registered on the "reference tag information" field 141 of the reference tag information management table 1214. The image data of the reference mark inputted newly is stored in the reference tag information storage means 1215 in order to apply it to the succeeding researching, meanwhile the pointer information to the mark image data is registered on the "pointer to reference mark image" field 142 of the reference tag information management table 1214. In addition, the characteristics value of the new reference mark on the Moment Invariants is calculated and then registered on the "attribute values (M1 to M6)" field 143 of the reference tag information management table 1214.

As described above, a user executes the input of the new mark image and the document tag information; thereby the new reference tag information can be registered.

Beside, in the above explanation, the reference tag information correlated with the reference mark is the character string applied to the reference mark as shown in Fig14 and Fig. 15(a), but the reference tag information needs not always to be the character string. That is to say, each reference mark can be correlated with any reference tag information in the reference tag management table 1214.

For example, instead of the reference tag information of the above-mentioned character string, the thumbnail image of reference mark is correlated with the reference mark as the reference tag information

respectively. The thumbnail image is printed on a researching sheet. By reading the thumbnail image of the researching sheet by a scanner, the desired document image can be searched.

In order to specify the document tag information appointing page among the entire inputted images, the two-dimensional code is applied as shown in the explanatory diagrams in Fig. 13 and Fig. 16. However the one-dimensional code and so on can be used, too. There are other methods for specifying the document tag information appointing page than the above, that is, a method that a specific mark is used instead of the two-dimensional code image, a method that a specific colored sheet is used, or a method that a specific formed sheet or a specific sized sheet is used. It is possible to obtain the same effect by those methods.

In addition, when the same document tag information is imparted to the entire inputted images, it is possible to arrange the document image processor by defining that only the image to be inputted as the first sheet is the image of the document tag information appointing page. In this case, since it has already been known that the image inputted as the first sheet is the document tag information appointing page, it needs not the processing for specifying the document tag appointing page by the two-dimensional code image and so on. Therefore, it is possible to simplify the processing as the whole.

It is needless to say that it is possible to extract the mark by searching the entire pages of the paper document without using the two-dimensional code. At this time, it will happens that characters included in the paper document, such as the "Confidential", and etc. are extracted as a

mark in addition to the mark attached by a user. In this case, the characters may be added to the mark management table 1213 as one of the mark.

The correlating between the mark image and the reference tag information is performed by using the characteristics value on the Moment Invariants in the above explanation, however it is possible to obtain the same effect by the correlating of the templates matching that compares the rate of the black pixels matched by overlapping two images.

Besides, it is possible to correlate a plurality of the reference marks with a piece of reference tag information. This is carried out by the following method; a plurality of the same reference tag information is registered on the reference tag information management table 1214, and then is correlated with the different reference mark respectively. In this case, after the paper document attached with different marks is inputted, the document images thus inputted is attached with the same document tag information.

Conversely, one reference mark can be correlated with a plurality of reference tag information. This is carried out by the method that the different reference tag information in the reference tag information management table 1214 is correlated with the same reference mark. In this case, after the paper document attached with one mark, the document image thus inputted is attached with a plurality of the document tag information.

(EMBODIMENT 6)

This embodiment describes the method for imparting the

document tag information to the document image by extracting the mark stamped at the blank part of the paper document to be registered. The followings express the points different from that of the embodiment 5 with reference to Fig. 12.

The image inputting means 1201 obtains document images by electronizing a paper document inputted by a user, like the embodiment 5. As shown in Fig. 17(b), the document tag information of "Confidential", "A-company", and "year 99" is attached to the document images 31 and 32, while the document tag information of "Confidential" and "B-company" is attached to the document image 33. In order to perform processing, the blank part of each image is stamped with a mark correlated with the document tag information to be attached with.

The image data obtained here is stored in the image memory 1202 temporarily. And after the data is compressed by the image data compressing means 1203, it is stored in the image storage area 1211 of the storage means 1210. As the information about the stored image data, the necessary information is registered respectively in the "image ID" field 121' and the "pointer to image data" field 122' of the registration image management table 1212' as shown in Fig. 17(a), which is the same as in the embodiment 5.

The image of the image memory 1202 is sent to the mark extracting means 1205' after the binarization by the image binarization means 1204. In order to extract the region of the mark image precisely, this embodiment uses a mark with frame as shown in Fig. 17(b) and performs the extracting of each mark by the mark extracting means 1205' as follows.

Each black pixel of the binary image is labeled, and regarding each of the black pixels contiguous components the size of the circumscribing rectangle is calculated. At this time, the size of the circumscribing rectangle of the black pixels contiguous components corresponding to the frame portion of the mark is large sufficiently comparing the each character size included in the inputted image, but does not get large extremely because the mark is stamped within the blank part of the document. By applying the characteristics, out of the black pixel contiguous components obtained by the labeling, only the region of which the circumscribing rectangle has the size between the specified two threshold values is extracted. That is to say, by extracting the region of the black pixels contiguous components wherein the respective sizes of the height and the width are larger than the specific threshold value (that is considered as the minimum size of the blank (the height and the width)), and less than another threshold value (that is considered as the maximum size of the blank), it is possible to extract the region of each mark image.

The number of marks extracted from the document images by the above processing is registered respectively on the "number of marks" field 124' of the registration image management table 1212'. Each extracted mark image is imparted with a mark ID respectively. The mark ID is registered in the "mark ID" field 131' of the mark management table 1213'. In addition, the information about the image ID of the inputted image attached with the mark, the information about the position that the mark was attached, and the information about the mark size are registered on the "image ID" field 132', the "position" field 134', and the "size" field 135' of

the mark management table 1213', respectively.

The embodiment of the invention is arranged so as to impart the document tag information to the image attached with the mark only. Beside, if a user wants to manage the images inputted between the image attached with the first mark and that with the next mark as a series of the document images included in the image with the first mark, the management method for imparting the mark management group No. can be adopted like the embodiment 5.

Like the embodiment 5, the calculating means 120A (the characteristics value calculating means 1206 and the similarity calculating means 1207) and the document tag information imparting means 1208 specify the document tag information correlated with the mark image according to the characteristics value of the Moment Invariants of the well-known technology. And the specified document tag information is registered on the "document tag information" field 133' of the mark management table 1213'.

If the above-mentioned processing is adopted, by the simple inputting that a mark is stamped on the blank of the paper document to be registered, it is possible to imparting the document tag information by the automatic searching. In this case, it is not necessary for the document tag information appointing page used in the embodiment 5, and the only document to be registered is inputted. As described above, the registration image management table 1212' and the mark management table 1213' are configured simply more than that of the registration image management table 1212 and the mark management table 1213 in the embodiment 5.

It is needless to say that this embodiment can adopt the method for imparting the two-dimensional code to the page attached with a mark in order to speed up the mark extracting.

In the invention of this embodiment, the mark stamped on the blank part of the side describing the content of the paper document is inputted. However, when a scanner and etc. can permit to scan both sides of the sheet, the input can be performed by stamping the mark on the backside. It can be also expected the same effect.

In addition, the mark has a frame, but the frame is not always necessary. In case of the mark without the frame, since it is considered generally that the mark be configured by the black pixels contiguous components of which size is larger than the characters included in the text of the paper document, it is possible to apply the embodiment.

As mentioned above, first of all, since the invention is configured that the region of which the region average character size is larger than the extracting criterion is extracted as the title region, it is possible to extract a plurality of title regions from one document image. In addition, it is possible to perform the extracting judgment on a plurality of levels according to the extracting parameters on a plurality of levels. Thereby the title region can be determined according to the characteristics of the document images inputted with the extracting parameter on a plurality of levels. Since the trim average method for calculating excluding both characters included in the specific rate of the larger side and those included in the specific rate of the smaller side is applied to the calculation of the total average character size and the calculation of the region average

character size, it is possible to improve the precision of the extracting.

Moreover, secondarily, the invention can impart the document tag information to the inputted image automatically by inputting the marked document to the document image processor without using the keyboard or the pointing device. By using the document tag information attached by the processing, the document image can be searched. Therefore, it is possible to manage and utilize the document image processor effectively.